Proceedings on Automation in Medical Engineering
Proc AUTOMED, 2026, Vol. 3, No 1, Article ID 2470, 2 Pages

Infinite Science Publishing

Training autoencoders on their own outputs causes collapse

Th. Schanze[1*]

[1] *IBMT, FB Life Science Engineering, Technische Hochschule Mittelhessen, Gießen, Germany*
[*] *Corresponding author, email: thomas.schanze@lse.thm.de*

Abstract: Classical autoencoders (AE) learn a compressed, meaningful representation of the input data and denoising autoencoders (DAE) capture the true underlying data manifold even when inputs are noisy. Data is the foundation of artificial intelligence, and thus for all autoencoder types. However, all types produce, when well trained, output data which are similar to the input data. This could lead to output data being added to the data that is to be used for further learning. We show on ECG signals that adding AE/DAE-generated reconstructions to the training set — intended to augment data — causes catastrophic performance collapse. This corroborates Shumailov's findings on 'model collapse' and the risks of data self-contamination.

I. Introduction

The classical autoencoder (AE) is a type of neural network (NN) that learns to efficiently compress and reconstruct data in an unsupervised manner. The goal is to reconstruct inputs as accurately as possible, whereby a bottleneck necessitates intelligent, compressed data representation [1].

Denoising autoencoders (DAE) are – by data and training design – forced to learn meaningful features instead of mapping identities [2]. They are heavily used in biomedical signal processing, especially in ECG denoising [3]. However, both types, i.e., AE and DAE, generate output data that is similar to the input data.

Data is the fundamental basis of artificial intelligence (AI), especially that of machine learning (ML) and thus of NN, because it provides the learning material. Sufficient, high-quality and, of course, large amounts of data still significantly determines performance of ML systems [4].

We will show, by using ECG signals, that the idea of adding the data generated by a AE or DAE during training to the training data set in order to virtually increase the amount of data, i.e., augmenting data for further learning, leads to a drastic drop in the performance of the AE or the DAE: the training data becomes increasingly similar, which also leads to increasingly similar responses. Ultimately this leads to network collapse. The following results support Shumailov's theory of "AI model collapse" and the risks of data self-contamination: quality declines as soon as AI models are trained with their own content [5], see also [6].

II. Material and methods

II.I. Denoising Autoencoder

Let $\mathbf{X} = [\mathbf{x}_1 \ \dots \ \mathbf{x}_M]$ be a "clean" set of M data vectors each of length D (the reference data). The AE is trained to reconstruct the clean input $\mathbf{x}_k$, while the DAE is trained to reconstruct the clean input from its corrupted version $\tilde{\mathbf{x}}_k = \mathbf{x}_k + \boldsymbol{\varepsilon}_k$: $\mathbf{y}_k = \mathbf{T}\{\tilde{\mathbf{x}}_k\} \cong \mathbf{x}_k$, where $\mathbf{T}$ represents AE's or DAE's signal transformation and $\boldsymbol{\varepsilon}_k$ is Gaussian white noise.

We selected fully connected denoising autoencoders with three layers and two different bottleneck sizes for study. The first architecture has layer sizes 95/20/95, and the second has 95/60/95. ReLU activations are used for all hidden layers, while the output layer is linear. We considered three different loss functions: mean squared error (MSE), mean absolute error (MAE), and structural similarity index measure (SSIM).

II.2. DAE training and recursive learning data manipulation of ECG signals

The starting point is classic AE/DAE training, which uses reference data for learning and validation, which belongs to the upper, i.e., gray, feedforward path of Fig. 1. After that a simple manipulation of the data for subsequent learning can start. Choose an arbitrary $j \in \{1, 2, \dots, M\}$ and let $\mathbf{y}_j$ be the response to $\mathbf{x}_j$, then the data for learning can be modified: $\mathbf{x}_j$ is to be replaced by $\mathbf{y}_j$, which relates to the data feedback path of Fig. 1. Thus we obtain the first modified data set $\mathbf{X}^{(1)} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{y}_j \ \dots \ \mathbf{x}_{M-1} \ \mathbf{x}_M]^{(1)}$. Then $\mathbf{X}^{(1)}$ is used for the next training step to adjust weights. This combined data manipulation and learning process is then continued until a stopping criteria is met. Note that this approach is intended to accelerate the effects of data recursion.

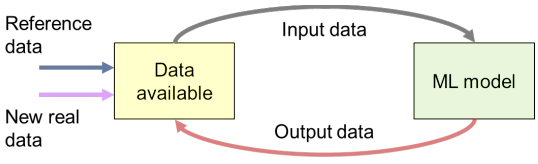

Figure 1: Schematic representation of a general model for data preparation for machine learning.

A data set of 300 ECG segments from 30 healthy patients, 10 of each patient, was used for training [7]. The number of

datapoints of an ECG segment containing P-wave, QRS-complex and T-wave is $D = 95$. All segments were QRS-aligned and scaled. The noise level for DAE was set to $\sigma(\boldsymbol{\varepsilon}_k) = 0.2\,\sigma(\mathbf{X})$. Network training was conducted using the Adam optimizer and a batch size of 300.

II.3. Measuring network collapse

The average pairwise correlation index between network model outputs is computed to measure network collapse of AE and DAE for every given parameter set. References are the average correlations of models without data feedback.

III. Results and discussion

We evaluated all six combinations of the two autoencoder types and three loss functions. For each combination, the networks were initialized randomly and a total of 480 simulations were performed, including 10 repetitions for each parameter class. Fig. 2 shows box-and-whisker plots of output correlations for AE and DAE with 20 hidden units each and MSE. Fig. 3 shows the results for 60 hidden units.

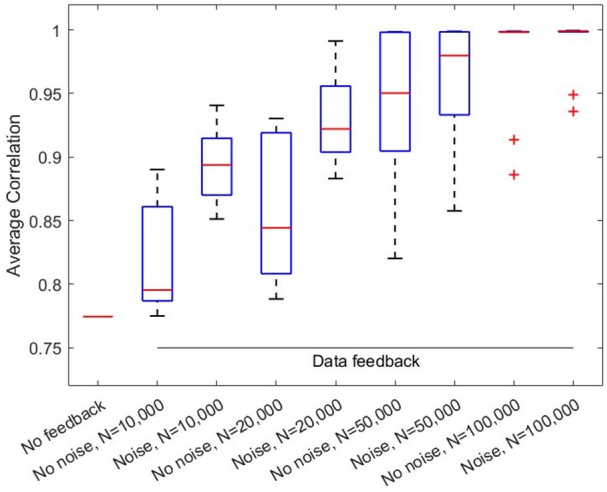

Figure 2: Box-and-whisker plots of the average pairwise correlation between the outputs of the network models as a function of the number of learning steps N and the network model, i.e., no noise: AE, noise: DAE. No feedback refers to the average pairwise correlation of the original data (almost identical here for the trained AE and DAE). The number of average correlations to create a box-and-whisker plot is 10 per class (8 data feedback classes), and the MSE loss function was used. The hidden layer has 20 neurons. Note, a correlation value of zero means no correlation, and a value of one means complete correlation.

Figures 2 and 3 clearly show that the correlation increases with the number of learning steps. After around 100,000 steps, the vectors for learning and output data are almost completely identical. In particular, the networks have lost their ability to provide adequate compression, reconstruction, and denoising properties for new data. This finding, which is shown in Figs. 1 and 2 for the MSE loss function, also applies to the other loss functions tested, i.e., MAE and SSIM. Thus a recursion of data, i.e., a training on self generated data, can lead to functional impairments of the network model, which is in accordance with Shumailov's theory of "AI model collapse" [5], see also [6]. This particularly means that also simple ML models can gradually or completely "forget" the actual underlying distribution of real data.

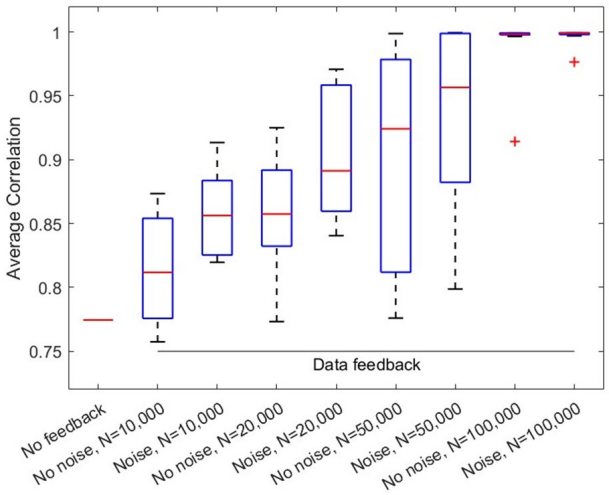

Figure 3: Same notation as in Fig. 2, but the hidden layer has 60 neurons.

Of course, the approach used for augmenting data for further network training is an exaggerated variant of a possible contamination of the input data for an ML system when the data created by the ML system is offered to the ML system for further learning. We may describe uncontrolled learning from self-generated data as a form of self-contamination.

Self-contamination or recursive data contamination, is already happening at scale on the open internet and may get much worse in the future [5]. According to Fig. 1, reference data, which can serve to check the proper functioning of an ML or AI system, and the quality of new real data are very important. Otherwise, diversity declines, results become more uniform, repetitive, and ultimately meaningless.

IV. Conclusions

The training data set is not only crucial but also represents the primary bottleneck in modern AI development. Preventing recursive contamination of training data — avoiding both data recursion and "digital inbreeding" — remains a critical open problem. Designating certain data sets as a "gold standard" could alleviate this issue.

Acknowledgments

We would like to thank Dr. Fars Samann and Max Bindemann for their valuable discussions.

Author's statement

The author declares no funding and no conflicts of interest.

References

[1] D. H. Ballard, *Modular learning in neural networks*, in Proceedings of the Sixth Nat. Conf. on Artificial Intelligence (AAAI-87), 1987.

[2] P. Vincent et al., *Extracting and composing robust features with denoising autoencoders*, in Proceedings of the 25th International Conf. on Machine Learning, ICML '08, pp. 1096–1103, 2008.

[3] F. Samann, T. Schanze, *AE-DD: Autoencoder-Driven Dictionary with Matching Pursuit for Joint ECG Denoising, Compression, and Morphology Decomposition*. AI, 6, pp. 234–250, 2025.

[4] A. Halevy et al., *The Unreasonable Effectiveness of Data*, IEEE Intelligent Systems, 24, pp. 8-12, 2009.

[5] I. Shumailov et al., *AI models collapse when trained on recursively generated data*, Nature, 631, pp. 755–759, 2024.

[6] T. Schanze, *Using signals generated by ECG denoising autoencoder during its learning results in a denoising performance collapse*, 58th Annual Meeting of the German Soc. of Biomed. Engn., 18 – 20 September 2024, Stuttgart, Biomed. Engn. / Biomed. Tech., 69, 2024.

[7] P. Wagner et al., *PTB-XL, A large publicly available electrocardiography dataset* (version 1.0.1), *PhysioNet,* 2020.